НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»
ФІЗИКО-ТЕХНІЧНИЙ ІНСТИТУТ
КАФЕДРА МАТЕМАТИЧНИХ МЕТОДІВ ЗАХИСТУ ІНФОРМАЦІЇ

«На правах рукопису»
УДК _____

«До захисту допущено»
В.о. завідувача кафедрою
_____ М.М.Савчук
(підпис) (ініціали, прізвище)
“___” _____ 2018р.

Магістерська дисертація

на здобуття ступеня магістра

зі спеціальності _____113 «Прикладна математика»_____
(код і назва)

на тему: _____Технології Big Data в інформаційній безпеці_____
_____

Виконав (-ла): студент (-ка) _6_ курсу, групи _ФІ-73мп_
(шифр групи)

_____Столова Олеся Вячеславівна_____ _____
(прізвище, ім’я, по батькові) (підпис)

Керівник _____професор кафедри ММЗІ, д.т.н., с.н.с. Кудін А.М._ _____
_____
(посада, науковий ступінь, вчене звання, прізвище та ініціали) (підпис)

Консультант _____ _____ _____
(назва розділу) (науковий ступінь, вчене звання, , прізвище, ініціали) (підпис)

Рецензент _____Проскуровський Р.В._____ _____
(посада, науковий ступінь, вчене звання, науковий ступінь, прізвище та ініціали) (підпис)

Засвідчую, що у цій магістерській дисертації немає запозичень з праць інших авторів без відповідних посилань.
Студент _____
(підпис)

Київ – 2018 року

Національний технічний університет України
«Київський політехнічний інститут
імені Ігоря Сікорського»
Фізико-технічний інститут
Кафедра математичних методів захисту інформації

Рівень вищої освіти: другий (магістерський) за освітньо–професійною програмою

Спеціальність: 113 «Прикладна математика»

ЗАТВЕРДЖУЮ
В.о. завідувача кафедрою
_____ М.М.Савчук
(підпис) (ініціали, прізвище)
«___»_____201_ р.

ЗАВДАННЯ
на магістерську дисертацію студенту
_____Столовій Олесі Вячеславівні_____
(прізвище, ім'я, по батькові)

1. Тема дисертації _____Технології Big Data в інформаційній безпеці_____
_____
_____ ,
науковий керівник дисертації професор кафедри ММЗІ, д.т.н., с.н.с. Кудін А.М. _____ ,
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

затверджені наказом по університету від _____р. № _____

2. Термін подання студентом дисертації _____

3. Об'єкт дослідження взаємозв'язки між логами _____
_____

4. Предмет дослідження (Вхідні дані – для магістерської дисертації за освітньо–професійною програмою)
журнали реєстрації подій _____
_____
_____

5. Перелік завдань, які потрібно розробити _____
_____
_____
_____
_____

6. Орієнтовний перелік ілюстративного матеріалу _____

7. Орієнтовний перелік публікацій _____

8. Консультанти розділів дисертації*

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв

9. Дата видачі завдання _____

Календарний план

№ з/п	Назва етапів виконання магістерської дисертації	Термін виконання етапів магістерської дисертації	Примітка
1	Вибір та затвердження теми магістерської дисертації	01.09-07.09	
2	Збір необхідної літератури	08.09-16.09	
3	Опрацювання теоретичного матеріалу	17.09-20.10	
4	Узгодження методів та інструментів для виконання поставленої задачі	21.10-31.10	
5	Реалізація практичної частини	1.11-22.11	
6	Аналіз отриманих результатів	23.11-30.12	
7	Оформлення магістерської дисертації	01.12-03.12	
8	Отримання допуску до захисту	04.12	

Студент _____ _____
(підпис) (ініціали, прізвище)

Науковий керівник дисертації _____ _____
(підпис) (ініціали, прізвище)

* Консультантом не може бути зазначено наукового керівника магістерської дисертації.

РЕФЕРАТ

Обсяг роботи 49 сторінок, 11 ілюстрацій, 2 таблиці, 17 джерел літератури.

Об’єкт досліджень – процес протоколювання подій інформаційних систем, що не призначені для обробки інформації з обмеженим доступом.

Предметом дослідження даної роботи є процес виявлення інформації з обмеженим доступом(персональних даних) в системах протоколювання подій.

Було зроблено аналіз методів технології Big Data, а також розглянуто правила та вимоги щодо обробки даних, зокрема персональних. Було визначено, що навіть виконуючи усі вимоги, щодо збору, обробки, зберігання та захисту даних, - в глобальних та корпоративних мережах присутні невраховані персональні дані. Розглянувши детально вміст лог-файлів інтернет-магазину – було виявлено досить багато інформації, якою можуть скористатися шахраї задля заподіяння шкоди користувачам цього ресурсу або його власникам. Серед таких даних наявні: імена користувачів, ip-адреси, браузери, ОС, пристрій та інше. Тож, використавши одну з технологій Big Data – Splunk – до журналів, було отримано наочні приклади «портретів користувачів», що містять інформацію, яка може бути використана злодіями для своїх цілей.

Отже, на основі проаналізованої інформації та отриманих результатів було розроблено методику виявлення неврахованих персональних даних в глобальних та корпоративних мережах.

BIG DATA, ІНФОРМАЦІЙНА БЕЗПЕКА, ОБРОБКА ЛОГІВ, БАЙЄСІВСЬКИЙ ПІДХІД, КЛАСТЕРНИЙ АНАЛІЗ, ЙМОВІРНІСТЬ, ПЕРСОНАЛЬНІ ДАНІ

РЕФЕРАТ

Объем работы 49 страниц, 11 иллюстраций, 2 таблицы, 17 источников литературы.

Объект исследований – процесс протоколирования событий информационных систем, непредназначенных для обработки информации с ограниченным доступом.

Предметом исследования данной работы является процесс выявления информации с ограниченным доступом (персональных данных) в системах протоколирования событий.

Был сделан анализ методов технологии Big Data, а также рассмотрены правила и требования по обработке данных, в частности персональных. Было определено, что даже выполняя все требования, по сбору, обработке, хранению и защите данных – в глобальных и корпоративных сетях присутствуют неучтенные персональные данные. Рассмотрев подробно содержание лог-файлов интернет-магазина – было обнаружено достаточно много информации, которой могут воспользоваться мошенники для причинения вреда пользователям этого ресурса или его владельцам. Среди таких данных имеются: имена пользователей, ip-адреса, браузеры, ОС, устройства и прочее. Поэтому, используя одну из технологий Big Data – Splunk – к журналам, было получено наглядные примеры «портретов пользователей», содержащих информацию, которая может быть использована мошенниками для своих целей.

Итак, на основе проанализированной информации и полученных результатов была разработана методика выявления неучтенных персональных данных в глобальных и корпоративных сетях.

BIG DATA, ИНФОРМАЦИОННАЯ БЕЗОПАСНОСТЬ, ОБРАБОТКА ЛОГОВ, БАЙЕСОВСКИЙ ПОДХОД, КЛАСТЕРНЫЙ АНАЛИЗ, ВЕРОЯТНОСТЬ, ПЕРСОНАЛЬНЫЕ ДАННЫЕ

ABSTRACT

Thesis consists of 48 pages, 11 illustrations, 2 tables, 17 literature sources.

The research object is the process of recording the information systems events that are not intended for processing of restricted access information. The subject of the following research is the process of identifying information with limited access (personal data) in event logging systems.

An analysis of several methods of Big Data technology was done, as well as the rules and requirements for personal data analysis were reviewed. It has been determined that even following all requirements for data collection, processing, storage and protection in global and corporative networks an unaccounted personal data is still present there. Having accurately examined the content of the log files from an online store, a lot of information was found that can be uses by fraudsters to harm users of this resource or its owners. Among these data are: user names, ip-addresses, browsers, operating systems, used devices, and other. Therefore, applying one of the Big Data technologies − Splunk − for log journals, illustrative examples of “user portraits” were obtained, containing information that can be used by fraudsters for their purposes. Therefore, based on the analyzed information and the obtained results, a method was developed for identifying unaccounted personal data in global and corporate networks.

BIG DATA, INFORMATION SECURITY, LOG PROCESSING, BAYES APPROACH, CLUSTER ANALYSIS, PROBABILITY, PERSONAL DATA

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ ТЕРМІНІВ

ІБ − інформаційна безпека

ІТ − інформаційні технології

ПЗ − програмне забезпечення

SPL − Search Processing Language

IBM − International Business Machines

GDPR − General Data Protection Regulation

OWASP − Open Web Application Security Project

ВСТУП

В даний час все більше уваги приділяється задачі ефективної обробки даних великих об’ємів, якими володіють державні служби та підприємства. Аналіз даних в «ручному режимі» часто є нездійсненним завданням, адже кількість інформації, що надходить до інформаційних систем, збільшується в геометричній прогресії. Ця задача ускладнюється, в тому числі за рахунок різноманітності типів пристроїв, з яких надходить інформація. Щоб людина змогла проаналізувати великі масиви даних, їх необхідно перевести в зрозумілий для неї вид; з декількох тисяч рядків слів і цифр отримати зрозумілу схему, діаграму або таблицю. Виходячи з цього можна сказати, що аналіз динамічно мінливої неструктурованої інформації в ручному режимі має велику ймовірність не помітити людським оком приховані взаємозв’язки між даними, в той час як використання коректно оброблених даних надає можливість вирішувати широкий спектр задач, а також передбачати та запобігати виникненню багатьох проблем.

Ще одна проблема, яка нерозривно пов’язана зі збільшенням об’ємів даних − зростання різноманітності злочинних атак на ці дані, серед яких особливу увагу звертають на себе персональні дані.

Для вирішення подібного типу завдань створено ряд технологій Big Data. Продуктами даного класу є IBM i2 та Splunk, що застосовуються в різних областях, включаючи інформаційну безпеку, управління IT інфраструктурою, аудит і моніторинг системних журналів. Дані рішення використовується в широкому спектрі галузей від охорони здоров'я, до фінансових послуг і промислового виробництва.

Актуальність роботи полягає в тому, що в зв’язку з набуттям чинності принципово нового законодавства про безпеку персональних даних, підвищується актуальність задачі управління персональними даними.

Мета роботи - на основі аналізу способів обробки даних великих об’ємів, а також правил обробки персональних даних, розробити методику

виявлення наявності в глобальних та корпоративних мережах неврахованих персональних даних.

Завдання роботи:

1. Розглянути принцип роботи систем Splunk та IBM i2.
2. Проаналізувати правила та вимоги встановлені ЄС щодо роботи із персональними даними.
3. Дослідити вміст журналів реєстрації подій.
4. Розробити методику виявлення наявності в комп’ютерних мережах неврахованих персональних даних.

Об’єкт дослідження: системи протоколювання подій інформаційних систем, що не призначені для обробки інформації з обмеженим доступом.

Предмет дослідження: процес виявлення інформації з обмеженим доступом(персональних даних) в системах протоколювання подій.

Наукова новизна одержаних результатів полягає в розробці методики виявлення наявності в глобальних та корпоративних мережах неврахованих персональних даних.

1 ТЕОРЕТИЧНІ ЗАСАДИ ВИВЛЕННЯ ПЕРСОНАЛЬНИХ ДАНИХ В ГЛОБАЛЬНИХ ТА КОРПОРАТИВНИХ МЕРЕЖАХ

1.1 Технології Big Data

В наш час є досить актуальною проблема забезпечення безпеки даних. Вирішення проблеми такого роду ускладнюється, через те, що дані зберігаються у великій кількості розрізнених систем з великим, постійно зростаючим їх об’ємом. Заволодіння чужою інформацією зазвичай відбувається по різних каналах (інтернет, особиста комунікація і т.д.), часто за участю інсайдерів. Розслідування таких ситуацій можливе тільки шляхом використання всієї необхідної інформації і її поданням у зрозумілому для спеціалістів форматі.

Для аналізу Big Data, на сьогодні, на ринку можна знайти багато рішень від різних компаній. Системи такого класу користуються все більшим попитом на ринку, адже вони дозволяють аналітикам проводити швидкий пошук закономірностей і прихованих зв'язків в даних, а також наочно представляти об'єкти аналізу. Що в свою чергу дає змогу вирішувати завдання широкого класу: прогнозування, прийняття рішень, координація ресурсів, поліпшення обслуговування, розслідування злочинів, аналіз ризиків, невизначеностей і загроз.

1.1.1Система Splunk та принцип її роботи

Splunk − це система для збору, зберігання, обробки та аналізу машинних даних, що представлені на усіх рівнях IT інфраструктури(фізичному, віртуальному та хмарному). Однією з головних особливостей є те, що вона збирає дані будь-якого формату з будь-яких джерел (включаючи події з системних журналів, дані про відвідування сайтів,

логи міжмережевих екранів і мережевих пристроїв, веб-серверів і баз даних(Рисунок 1.1)), і тому список можливих застосувань системи дуже широкий. Дана характеристика є дуже корисною, адже виключає необхідність розробки спеціалізованих парсерів/конекторів, їх підтримки та модифікації у випадку зміни початкових форматів.

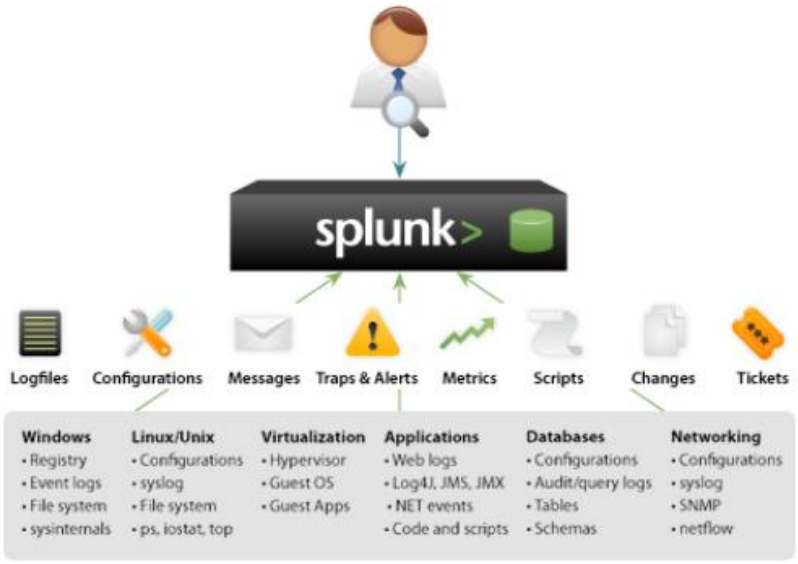

Рисунок 1.1. Джерела надходження даних

Принцип роботи даної системи виглядає таким чином: є робочі машини, що створюють і передають на сервер Splunk логи, який в свою чергу зберігає, індексує і дозволяє аналізувати їх. Splunk здійснює збір, пошук, моніторинг та аналіз даних в режимі реального часу на надзвичайно великих обсягах даних.

Система Splunk використовує технологію MapReduce, що забезпечує розподіл навантажень і горизонтальну масштабованість системи, і дозволяє швидко обробляти дуже великі обсяги даних. При збільшенні даних, сервер Splunk можна зробити кластером з декількох фізичних машин (між якими буде розподілятися зберігання інформації і які будуть використовуватися для її обробки) і таким чином розподілити навантаження [1].

Існує багато способів передавати логи з робочих машин: можна власноруч відсилати дані в Splunk по TCP / IP (наприклад, замість того, щоб писати в файл), або ж через спеціальну програму(forwarder), яка вміє швидко

і ефективно відсилати зміни логів на сервер, через такі технології як NFS/SMB, або SNMP). Під ОС Windows Splunk вміє брати дані з реєстру, Windows Events.

Splunk сприймає логи як текстову інформацію, що є розбитою на рядки. За рахунок процесу індексування вхідні дані розбиваються на поля і значення.

Далі за допомогою спеціальної мови запитів SPL, можна працювати з цими полями: фільтрувати, агрегувати, сортувати, складати звіти, формувати таблиці, звертатися як до внутрішніх, так і зовнішніх довідників, створювати dashboard'и, з широким спектром візуалізації. Задавши порогові значення за певним параметром можна формувати оповіщення про інциденти або ж створювати правила і система сама буде реагувати на них за допомогою запуску скриптів.

Логи, збережені в системі за увесь час, є доступними для запитів, тобто архівування не виконується. Така функція наявна, за рахунок, використання технології Hadoop Data Roll, що забезпечує зниження витрат на зберігання історичних даних (іноді до 80%). Можна без витрат переносити дані Splunk в озеро даних Hadoop зі збереженням усіх можливостей пошуку [2].

Також в Splunk можна здійснювати пошук по даним, що накопичені протягом всього часовому проміжку, за який вони були зібрані. Тобто пітримується функція пошуку, моніторингу, оповіщення, звітності і аналізу за будь-який час (історичні дані і дані в реальному часі в одному рішенні). В Splunk, не має потреби заздалегідь знати структуру даних, щоб сформувати запит, оскільки пошук здійснюється за часом. Це підвищує гнучкість системи. Ще наявні зручні функії такі як зупинка будь-якого запиту, пауза та представлення проміжних результатів.

Створюючи панелі (dashboards), можна формувати власний Splunk-додаток. У Splunk є магазин додатків де наявні багато готових конфігурацій для аналізу популярних систем, наприклад, UNIX syslog, логи Apache,

Microsoft Exchange і т.д. Splunk підтримує як Windows / Linux, так і OS X, FreeBSD, а крім того, і HP-UX та AIX [2].

Отже, Splunk є універсальною платформою, що об'єднує дані та надає широкі можливості з їх аналізу, побудови звітів та візуалізації. Завдяки таким можливостям даного рішення, можна легко і в автоматичному режимі виявляти атаки, приймати зважені рішення та прогнозувати імовірні ситуації.

1.1.2 Огляд системи IBM I2

IBM i2 - це передове рішення в області розслідувань економічних, юридичних та IT-злочинів. Головне завдання IBM i2 - пошук прихованих взаємозв'язків та закономірностей серед великої кількості сутностей, що дозволяє використовувати його службам економічної та внутрішньої безпеки, ризик-менеджерами і розслідувальним управлінням. i2 працює на стику трьох напрямків: боротьби з шахрайством, IT і економічної безпеки, і робить процес аналізу простим і наочним. Вирішення завдання безпеки інформації реалізується за допомогою візуалізації прихованих зв'язків і статистичних закономірностей між даними. IBM i2 дозволяє представити результати обробки в зручній формі, роблячи їх презентабельними і наповненими вичерпною інформацією[3].

Система IBM i2 складається з різних компонентів, які, в залежності від розв'язуваних завдань, можуть бути налаштовані і застосовані в різних комбінаціях(Рисунок1.2)

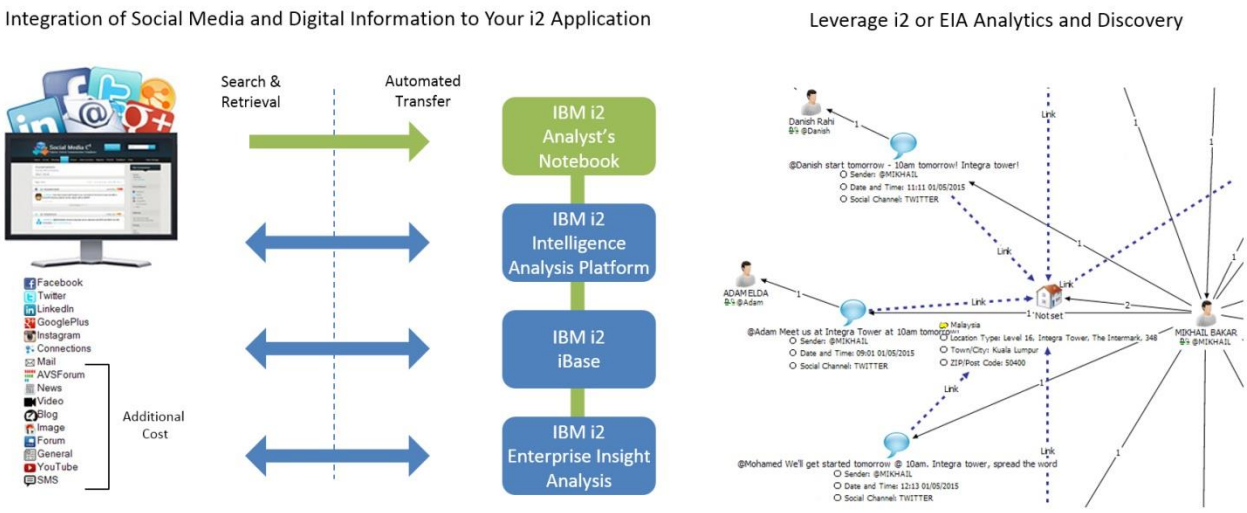

Рисунок 1.2 Принцип роботи IBM i2

1) Analyst's Notebook − візуальне аналітичне середовище, яке дозволяє максимально ефективно використовувати величезні обсяги інформації, накопичені компаніями та установами. Analyst's Notebook, беручи логи з цілого ряду джерел,− перетворює їх на цінну інформацію, що дає змогу побудувати повну картину для дослідження. У Analyst's Notebook дані зберігаються як об'єкти, зв'язки та властивості. Завдяки простому та зрозумілому інтерфейсу, аналітики можуть швидко зіставляти, аналізувати і наочно представляти дані, скорочуючи час на пошук важливої інформації в неструктурованих даних. Забезпечує тимчасовий і геопросторовий аналіз завдяки інтеграції з геопросторовими функціями ArcGIS Server компанії Esri.

Дане рішення надає актуальні і дієві аналітичні засоби, що допомагають виявляти, передбачати, запобігати і припиняти злочинну, терористичну і шахрайську діяльність.

Тож, серед задач, що вирішує IBM i2 Analyst's Notebook можна виділити такі:

- швидка систематизація розрізнених даних і подання в єдиному узгодженому вигляді;
- визначення ключових осіб, подій, зв'язків і закономірностей, які не завжди можна виявити іншими засобами;
- покращене розуміння структури, ієрархії і способів дій злочинних, терористичних і шахрайських організацій;

- спрощення обміну складними даними, що дозволяє приймати своєчасні і точні оперативні рішення.

2) iBase – компонент, що дозволяє збирати й аналізувати дані з різних джерел. iBase є центральним вузлом системи, за допомогою якого здійснюється зберігання і аналіз інформації. Також за допомогою цього компонента здійснюється управління обліковими записами і правами доступу для багатокористувацьких систем.

iBase дозволяє:

- спроектувати базу даних, в якій буде згодом зберігатися інформація;
- формувати правила імпорту даних зі сторонніх джерел;
- здійснювати контроль доступу до даних і аудит подій такого доступу;
- централізовано зберігати результати аналізу.

3) iBridge - підключається до обраних корпоративних баз даних з можливостями пошуку/запиту для повернення даних, готових до аналізу.

iBridge дозволяє:

- безпосередньо підключатися до зовнішніх баз даних.
- здійснювати пошук елементів в базі даних.
- формувати візуальні запити на пошук інформації[4].

Також, існують такі компонени лінійки i2 як:

- i2 Pattern Tracer, що оперативно аналізує великі обсяги телефонних записів, групує їх за спільними ознаками і виявляє ключових учасників. Додаток швидко ідентифікує потенційні цілі дзвінків і допомагає запобігти майбутнім інциденти.
- i2 Chart Reader дозволяє аналітикам обмінюватися результатами аналізу з тими, у кого немає доступу до i2 Analyst's Notebook.
- i2 Text Chart: інтуїтивно зрозуміле, кероване користувачем вилучення тексту та його візуалізація допомагає аналізувати неструктуровані дані.
- i2 Text Chart Auto Mark: автономний додаток, що автоматично виявляє і виділяє об'єкти, що вас цікавлять, в різноманітних документах.

Отже, система даного класу, дозволяє отримувати реальну віддачу від IT даних шляхом надання операційної аналітики, яка згодом може бути використана для поліпшення рівня обслуговування користувачів, скорочення експлуатаційних витрат, зниження ризиків у інформаційній безпеці, створення нових продуктів і послуг, а також здатна протидіяти кіберзлочинності, розслідувати факти неправомірного доступу до інформації.

1.2 Правила роботи з персональними даними

Збір, аналіз та переміщення персональних даних по всьому світу набули великого економічного значення.

Персональні дані - це будь-яка інформація, що стосується конкретної фізичної особи, за допомогою якої можна її ідентифікувати. Сюди відносяться як відомості, які користувач самостійно надав для того чи іншого веб-ресурсу (ім'я та прізвище, стать, email або номер телефону), так і дані, зібрані автоматично. Наприклад, це інформація про місцезнаходження, про пристрій, операційну систему, логін користувача web-ресурсу і т.д.

Крім того, персональними даними в інтернеті вважають відомості веб-перегляду, пошукові запити, пости в соціальних мережах та всю ту інформацію, на основі якої можна визначити інтереси користувача, його соціальний статус, релігійні переконання, політичні погляди і т.п. В окрему групу можна виділити платіжні дані, які, безумовно, також належать до персональних.

Оскільки персональні дані є дуже важливою складовою сучасного світу, а також причиною багатьох злочинів, були створені правила та вимоги збору, зберігання, передавання та захисту таких даних.

1.2.1 Загальні правила захисту даних(GDPR)

GDPR (General Data Protection Regulation, або загальні правила захисту даних) , надає резидентам ЄС інструменти для повного контролю над своїми персональними даними.

Підхід до захисту персональних даних в GDPR заснований на восьми принципах, які були задокументовані ще в 1980 році (в «Керівництві про захист приватного життя та транскордонних потоків персональних даних») і схвалені як ЄС, так і США. У GDPR вони відображені в семи пунктах:

1. Принцип законності, справедливості і прозорості. Персональні дані повинні бути отримані законними і справедливими засобами за згодою суб'єкта даних.
2. Конкретизація мети. Мета збору даних повинна бути вказана під час збору, і дані не повинні використовуватися ні для чого іншого, крім початкового наміру.
3. Мінімізація даних. Зібрані дані повинні відповідати заданій спочатку цілі. Забороняється збирати дані в більшому обсязі, ніж це потрібно для досягнення мети.
4. Точність. Персональна інформація повинна бути точною, повною та актуальною, наскільки це необхідно для заданих цілей. Якщо такі дані будуть вважатися неточними, вони повинні бути видалені або виправлені (на вимогу користувача).
5. Обмеження зберігання. Дані зберігаються в формі, яка дозволяє ідентифікувати користувача не довше, ніж це необхідно для виконання цілей обробки інформації.
6. Цілісність і конфіденційність. Особисті дані повинні бути захищені гарантіями безпеки від таких ризиків, як втрата або несанкціонований доступ, знищення, використання, модифікація або розкриття даних.
7. Підзвітність. Контролер несе відповідальність і повинен бути готовий продемонструвати дотримання заходів, зазначених вище[4].

Важливим моментом є те, що GDPR застосовний до обох – як до оброблювача даних так і контролера даних. Контролер даних визначає мету і значення обробки персональних даних, а обробник відповідальний за безпосередню обробку даних, але обидва несуть відповідальність за дотримання норм GDPR.

Регламент GDPR замінив Директиву про захист даних від 1995 року. Постанову було прийнято 14 квітня 2016 року, почав застосовуватися з 25 травня 2018 року. В ньому розширено поняття персональних даних, введено поняття «транскордонної передачі даних», «псевдонімізації», встановлено «право на забуття», визначена роль посадової особи щодо захисту даних.

Були введені поняття: контролера даних - організація, яка збирає дані від резидентів ЄС; обробника даних − організація, яка обробляє дані від імені контролера даних, наприклад, постачальник хмарних послуг; суб'єкта даних (особа) − фізична особа; спеціальна категорія персональних даних − дані про расу, політичній думці, релігійних або філософських переконаннях, генетичні дані, членство в профспілках, біометричні дані дозволяють визначити конкретну людину, дані про здоров'я, сексуальна орієнтація.

Дуже цікавим нововеденням є «право бути забутим», за яким стоїть право суб'єкта на стирання своїх персональних даних, яке повинен здійснити контролер без будь-якої безпідставної затримки, а також контролер забов’язаний видалити персональні дані без будь-якої необґрунтованої затримки у разі виникнення однієї наведених нижче підстав:

(a) немає більше потреби в персональних даних, для цілей, для яких їх збирали чи іншим чином опрацьовували;

(b) суб’єкт даних відкликає згоду, на якій ґрунтується опрацювання, згідно з пунктом (a) статті 6(1) чи пунктом (a) статті 9(2), та якщо немає іншої законної підстави для опрацювання:

(c) суб’єкт даних заперечує проти опрацювання згідно із статтею 21(1), та немає жодних першочергових законних підстав для опрацювання, або суб’єкт даних заперечує проти опрацювання згідно статтею 21(2);

(d) персональні дані опрацьовували незаконно;

(e) персональні дані необхідно стерти для дотримання встановленого законом збов’язання, закріпленого в законодавстві Союзу або держави-члена, яке поширюється на контролера;

(f) персональні дані збирали в зв’язку з пропонуванням послуг інформаційного суспільства, вказаних у статті 8(1) [6].

Дана стаття вище названого регламенту є досить важливою, адже дані суб'єкта, що були надані у свій час для благих цілей та не видалені після їх реалізацій (або недостатньо захищені), потрапивши у руки шахраїв можуть стати способом наживи для них.

Та все ж таки, навіть виконуючи усі вимоги GDPR, в комп’ютерних мережах можна виявити невраховані персональні дані, зібравши і обробивши які, можна заподіяти шкоди суб’єкту. Наприклад, якщо суб’єкт користується Інтернет-магазинами,− він надає багато персональної інформації(ім’я, адресу, дату народження, email, іноді паспортні дані, номери карток). При цьому немало інформації можна дізнатися з інформації веб-сеансу(дані про браузер, IP-адресу, геолокацію, вигляд пристрою, операційну систему та інше). Нижче розглянемо, як же забезпечується безпека web-додатків та побачимо, що саме злочинці можуть використати для своїх шахрайських намірів.

1.2.2 Проект забезпечення безпеки web-додатків

Open Web Application Security Project (OWASP) − відкритий проект забезпечення безпеки web-додатків. Дана організація опублікувала документ «Top 10 Proactive Controls for Software developers» («Топ-10 проактивних інструментів управління для розробників програмного забезпечення»), що описує найбільш критичні аспекти, на яких повинні зосередитися розробники

ПЗ. Документ містить список технік по забезпеченню безпеки, обов'язкових для реалізації при розробці.

Список «Top 10 Proactive Controls for Software developers» починається з опису вимог безпеки, прописаних в промислових стандартах і законодавстві. Топ-10 вимог безпеки за версією OWASP виглядає наступним чином (в порядку важливості):

C1: Визначення вимог безпеки;

C2: Реалізація фреймворків і бібліотек безпеки;

C3: Захист доступу до баз даних;

C4: Шифрування і забезпечення безпеки даних;

C5: Перевірка автентичності всіх вхідних даних;

C6: Реалізація Digital Identity;

C7: Забезпечення управління доступом;

C8: Захист всіх даних;

C9: Реалізація авторизації та моніторингу;

C10: Обробка всіх помилок і виключень [7].

В даному документі міститься інформація щодо забезпечення безпечного трансферу даних, в тому числі персональних, а також їх шифрування.

Розглянувши способи збору даних, зберігання та передачі, можна побачити, що все ж таки існують лазівки, за допомогою яких злодії можуть дістатися до персональних даних. Наприклад, якщо лог інтернет-магазину містить user_name суб’єкта, то шахраї можуть методом підбору знайти правильний пароль та авторизуватися у особистий кабінет користувача. Або ж з IP-адреси, що міститься у логах, за допомогою одного з сервісів, можна дізнатися місце розташування користувача до територіальної одиниці – район [8].

Висновки до розділу 1

Для захисту великих даних сучасним підприємствам потрібні рішення, що здатні адаптуватися до нових загроз і мінливих бізнес-вимог. Щоб бути готовими до зовнішніх атак, діям внутрішніх зловмисників і шахрайським операціям, що призводять до великих витрат,− необхідний постійний моніторинг безпеки та відповідність вимогам, швидке реагування на інциденти і можливість виявляти і усувати відомі, невідомі і все більш витончені загрози.

В даному розділі було досліджено декілька систем, які завдяки потужним інструментам обробки даних, здатні вирішувати такі проблеми. Вони дозволяють швидко шукати, зіставляти, аналізувати і наочно представляти дані з різних джерел, швидко виявляти ключову інформацію серед складних даних, ефективно проводити аналіз системи взаємопов'язаних об'єктів і динаміки послідовних подій, відображаючи результати дослідження у вигляді зручних для розуміння схем і діаграм. Дані продукти допомагають швидко виявити джерело проблеми і визначити правильний напрямок подальшої роботи, що є дуже важливою якістю в боротьбі зі злочинами, пов’язаними з заволодінням цінною інформацією.

Окрім цього було розглянуто правила та вимоги, яких повинні дотримуватися компанії та організації, які збирають, обробляють та зберігають дані користувачів, а також було звернуто увагу на необхідність забезпечення безпеки даних з боку розробників ПЗ. Було наведено декілька прикладів того, яким чином та яку саме інформацію можуть отримати шахраї з логів задля вчинення своїх злочинів.

2 МЕТОДИ ОБРОБКИ ТА АНАЛІЗУ ВЕЛИКИХ МАСИВІВ ДАНИХ

Сучасні системи обробки даних використовують різноманітні методи та алгоритми для аналізу Big Data. Загальною рисою всіх методів є те, що вони дозволяють зручо та швидко опрацьовувати інформацію великого об'єму. В залежності від поставленої задачі підбираються методи, що найбільш оптимально її вирішують. В даному випадку, для аналізу вмісту журналів протоколювання подій в глобальних та корпоративних мережах з метою виявлення інформації з обмеженим доступом (персональних даних), було розглянуто Байєсівський підхід із використанням методу Монте-Карло та кластерний аналіз,а саме алгоритм k-means.

В даному розділі розглянуто теоретичну основу вищезгаданих методів. Вони є практично реалізованими у системі Splunk, робота з якою буде представлена у розділі 3 для виявлення інформації з обмеженим доступом(персональних даних).

2.1 Байєсівський підхід

Припустимо, що ми намагаємося дослідити деякі явища. Маємо деякі знання, отримані до спостережень/експериментів. Це може бути досвід минулих спостережень, модельні гіпотези, очікування. В процесі спостережень ці знання піддаються поступовому уточненню. Після спостережень/експериментів у нас формуються нові знання про явище. Будемо вважати, що ми намагаємось оцінити невідомі значення величини θ за допомогою спостережень деяких її опосередкованих характеристик $x|\theta$.

Відома формула Байеса встановлює правила, за якими виконується перетворення знань у процесі спостережень. Позначимо апріорні знання про величину θ за $p(\theta)$. В процесі спостережень отримуємо серію значень

$x = (x_1, ..., x_n)$. При різних θ спостереження вибірки x визначається значенням правдоподібності $p(x|\theta)$. За рахунок спостережень, уявлення про значення θ змінюються згідно з формулою Байеса:

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)d\theta} .$$

Зазначимо, що знаменник не залежить від θ та потрібен лише для нормування апостеріорної щільності [9].

Тож, перейдемо до Байєсівського підходу, ідея якого полягаєв переході від апріорних знань(або точніше незнання) до апостеріорних з урахуванням явищ, що спостерігаються.

Особливість цього підходу в тому, що по-перше усі величини та параметри вважаються випадковими, адже якщо значення параметрів розподілу невідоме, то вони випадкові з точки зору нашого незнання. По-друге, Байєсівські методи працюють навіть при об’ємі виборки 0! В цьому випадку апостеріорний розподіл дорівнює апріорному. В якості оцінок невідомих параметрів виступають апостеріорні розподіли, тобто вирішити задачу оцінювання деякої величини, означає знайти її апостеріорний розподіл.

Байєсівський підхід має свої недоліки. Починаючи з 1930х рр. Байєсовські методи піддавалися різкій критиці та практично не використовувались через наступні причини:

- в Байєсівських методах передбачається, що апріорний розподіл відомий до початку спостережень та не пропонується конструктивних способів його вибору;

- прийняття рішень при використанні Байєсівських методів у нетривіальних випадках потребує колосальних обчислювальних витрат, пов’язаних з чисельним інтегруванням в багатовимірних просторах;

- Фішером була показана оптимальність методу максимальної правдоподібності, а звідси – відсутність сенсу спроб вигадати щось краще.

З початку 1990 рр спостерігається відродження Байєсівських методів, які виявилися в змозі вирішити багато серйозних проблем статистики та машинного навчання.

При використанні методу Байєса можна обчислити точкові оцінки:

Математичне очікування за апостеріорним розподілом є досить трудомісткою функцією:

$$\hat{\theta}_B = \int \theta_p(\theta|x)d\theta$$

Максимум апостеріорної щільності зручний з точки зору обчислення:

$$\hat{\theta}_{MP} = \arg\max P(\theta|x) = \arg\max P(x|\theta)P(\theta) = \arg\max(\log P(x|\theta) + \log P(\theta))$$

Це фактично регуляризація методу максимальної правдоподібності.

Байєсівський вивід можна розглядати як узагальнення класичної булевої логіки. Тільки замість понять «істина» та «хиба» вводиться «істина з ймовірністю p».

Узагальнення класичного правила Modus Ponens: $\dfrac{A, A \Rightarrow B}{A \& B}$

$$\frac{p(A), p(B|A)}{p(A \& B)}$$

Якщо розглянемо ситуацію : $\dfrac{A \Rightarrow B, B}{A = ?}$

$$\frac{p(B|A), p(B), p(A)}{p(A|B)}$$

Формула Байєса дозволяє розрахувати зміни степеня істинності A з урахуванням про B. Це новий підхід до синтезу експертних систем. На відміну від нечіткої логіки, він теоретично обґрунтований та математично коректний [10].

2.1.1 Методи Монте Карло в Байєсівському підході

Розглянемо ймовірнісний розподіл $p(T)$. Методи Монте Карло (методи статистичних випробувань) передбачають генерацію виборки з цього розподілу: $T_1,...,T_N \sim p(T)$.

Ця вибірка може бути використана для оцінки ймовірнісних інтегралів виду:

$$\mathbb{E}_T f(T) = \int f(T) p(T) dT \simeq \frac{1}{N} \sum_{n=1}^{N} f(T_n) \quad (1)$$

До інтегралів такого виду зводиться багато кроків при здійсненні байєсівського виводу в ймовірнісних моделях. Наприклад, такими інтегралами є функції обгрунтованності $p(t|X,\alpha)$ та прогнозний розподіл $p(t_{new}|x_{new},t,X,\alpha)$, в моделі RVM, а також функціонал на М-кроці ЕМ-алгоритму $\mathbb{E}_{q(T)} \log p(X,T|\theta)$. Тут варто відзначити, що щільність $p(T)$ у подібних ймовірнісних інтегралах часто відома з точністю до нормуючої константи $p(T) = \frac{1}{Z} \tilde{p}(T)$.

Вибірка $T_1,...,T_N \sim p(T)$ може бути використана для оцінки моди розподілу $p(T)$: $\max_T p(T) \simeq \max_n p(T_n)$, так як поява точок вибірки найбільш ймовірне в областях великих значень щільності.

Основне питання, що розкривається в подальшому, полягає в тому, як ефективно згенерувати вибірку $T_1,...,T_N$ з ймовірнісного розподілу, заданого своєю щільністю $p(T)$ або своєю ненормованою щільністю $\tilde{p}(T)$ [11].

2.1.2 Найпростіші методи генерації

Розглянемо одновимірну випадкову величину X та її функцію розподілу $f(x) = \mathbb{P}\{X < x\}$.

Розглянемо випадкову величину $f(X)$, що має рівномірний розподіл на інтервалі $[0,1]$. Найпростіший спосіб генерації випадкової величини, заданої своєю функцією розподілу: спочатку генеруємо $\xi \sim R[0,1]$, а потім обчислюємо $x = f^{-1}(\xi)$. Даний метод генерації отримав назву методу оберненої функції. За допомогою цього методу можна згенерувати вибірку з довільного дискретного розподілу з кінцевим носієм.

Метод оберненої функції можна застосовувати тільки в обмеженій кількості випадків, так як він потребує аналітичного обчислення оберненої функції до функції розподілу. Зокрема, метод оберненої функції не можна застосувати для нормального розподілу.

Для генерації вибірки з нормального розподілу можна скористатися центральною граничною теоремою.

Одним з загальних методів генерації, який може бути застосований практично для будь-якої одновимірної неперервної випадкової величини, є метод Rejection sampling.

2.1.3 Ідея MCMC

Розглянемо тепер питання генерації вибірки з розподілу $p(T)$ в багатовимірному просторі за допомогою методів Монте Карло за схемою марківського ланцюга (MCMC). У цих методах вводиться деякий марківський ланцюг з апріорним розподілом $p_0(T)$ і ймовірностями переходу в момент часу n $q_n(T_{n+1}|T_n)$, а генерація вибірки виконується наступним чином:

$$\begin{aligned} &T_1 \sim p_0(T), \\ &T_2 \sim q_1(T_2|T_1), \\ &\vdots \\ &T_N \sim q_{N-1}(T_N|T_{N-1}). \end{aligned} \qquad (2)$$

Зауважимо, що при такому підході генерується вибірка, що не є набором незалежних випадкових величин. Однак, вона підходить для оцінки ймовірнісних інтегралів виду (1) або оцінки моди розподілу. У тому випадку, якщо необхідно отримати набір незалежних величин, досить прорідити отриманий набір $T_1,...,T_N$, взявши кожен m -ий відлік, де m досить велике.

Надалі розглядається питання про те, як вибрати ймовірності переходу $q_n(T_{n+1}|T_n) = q(T_{n+1}|T_n)$, таким чином, щоб вибірка, що генерується за схемою (2), була б вибіркою з розподілу $p(T)$ [12].

2.1.4 Схема Метрополіса-Гастингса

Нехай потрібно згенерувати вибірку з розподілу $p(T)$, що відомий нам з точністю до константи нормування: $p(T) = \frac{1}{Z}\tilde{p}(T)$.

Розглянемо крок генерації за схемою Метрополіса-Гастингса. Нехай на кроці n згенерована конфігурація T_n. Тоді на кроці $n+1$ спочатку генерується конфігурація T_* з деякого апріорного розподілу $r(T|T_n)$. Потім обчислюється величина $A(T_*, T_n) = \min(1, \frac{\tilde{p}(T_*)r(T_n|T_*)}{\tilde{p}(T_n)r(T_*|T_n)}$ та точка T_* береться в якості наступної точки T_{n+1} з ймовірністю $A(T_*, T_n)$. В іншому випадку $T_{n+1} = T_n$. Таким чином, ввели марківський ланцюг з ймовірністю переходу:

$$q(T_{n+1}|T_n) = \begin{cases} r(T_{n+1}|T_n)\,A(T_{n+1}, T_n), \text{якщо } T_{n+1} \neq T_n \\ 1 - r(T_{n+1}|T_n)\,A(T_{n+1}, T_n), \text{якщо } T_{n+1} = T \end{cases} \quad (3)$$

Покажемо, що розподіл $p(T)$ є інваріантним відносно введеного марківського ланцюга. Якщо $T_{n+1} = T_n$, то інваріантність зберігається, так як значення T_n не змінюється. Для випадку $T_{n+1} \neq T_n$ перевіримо можливість виконання рівняння детального балансу:

$$p(T_n)q(T\,|\,T_n) = \min(p(T_n)r(T\,|\,T_n), p(T)r(T_n\,|\,T)) = \min(p(T)r(T_n\,|\,T), p(T_n)r(T\,|\,T_n)) = p(T)q(T_n\,|\,T)$$

Для ергодичності введеного марківського ланцюга достатньо вимагати виконання $r(T\,|\,S) > 0, \forall T, S$.

В тому випадку, якщо апріорний розподіл є симетричним, тобто $r(T\,|\,S) = r(S\,|\,T), \forall S, T$, то схема Метрополіса-Гастингса переходить в класичну схему Метрополіса. Згідно з цією схемою, якщо значення щільності в новій точці T_* виявилося вище, ніж значення щільності в попередній точці T_n, то ця точка гарантовано приймається в якості наступної точки вибірки. Якщо щільність в новій точці виявилась меншою, то така точка може бути прийнята, але з ймовірністю, пропорційною величині зменшення щільності [12].

2.1.5 Схема Гіббса

Нехай необхідно згенерувати вибірку з багатовимірного розподілу $p(T)$, де $T = \{t_1, ..., t_p\}$. Розглянемо крок генерації за схемою Гіббса. Нехай на кроці n згенерована конфігурація $T^n = \{t_1^n, ..., t_p^n\}$. Тоді генерація наступної точки вибірки T^{n+1} виконується наступним чином:

$$\begin{aligned}
&t_1^{n+1} \sim p(t_1\,|\,t_2^n, t_3^n, ..., t_p^n),\\
&t_2^{n+1} \sim p(t_2\,|\,t_1^{n+1}, t_3^n, t_4^n ..., t_p^n),\\
&t_3^{n+1} \sim p(t_3\,|\,t_1^{n+1}, t_2^{n+1}, t_4^n, ..., t_p^n), \qquad (3)\\
&\dots\\
&t_p^{n+1} \sim p(t_p\,|\,t_1^{n+1}, t_2^{n+1}, ..., t_{p-1}^{n+1}).
\end{aligned}$$

Тут через $p(t_i\,|\,T_{\setminus i})$ позначено маргінальний одновимірний розподіл значень і-ої компоненти за умови усіх інших. Таким чином, згідно зі схемою Гіббса генерація вибірки з багатовимірного розподілу змінюється на ітеративну генерацію точок з одновимірного розподілу. За аналогією з методами одновимірної оптимізації генерація вибірки з одновимірного

розподілу є істотно легшою задачею, аніж генерація вибірки з багатовимірного розподілу.

Доведемо, що розподіл $p(T)$ є інваріантним відносно введеного марківського ланцюга. Розглянемо один крок генерації чергової компоненти $t_p \sim p(t_p \left| T_{\setminus p} \right.)$. За припущенням індукції $T_{\setminus p} \sim p(T_{\setminus p})$. Тоді сумісна конфігурація $(t_p, T_{\setminus p}) \sim p(t_p \left| T_{\setminus p} \right.) p(T_{\setminus p}) = p(T)$. Звідси, сумісний розподіл є інваріантним відносно одного кроку процесу генерації (3). Отже, він є інваріантним і відносно усього процесу (3).

При реалізації схеми Гіббса на практиці часто допускається наступна помилка: замість кроку $t_p^{n+1} \sim p(t_p \left| t_1^{n+1}, ..., t_{p-1}^{n+1}, t_{p+1}^{n+1}, ..., t_p^n \right.)$ виконується крок $t_p^{n+1} \sim p(t_p \left| t_1^n, ..., t_{p-1}^n, t_{p+1}^n, ..., t_p^n \right.)$, тобто в умові підставляються значення компонент тільки з попередньої ітерації. За такого підходу ймовірність переходу в марківському ланцюзі визначається як $q(T \left| T^n \right.) = \prod_{p=1}^{P} p(t_p \left| T_{\setminus p}^n \right.)$ (4).

Розподіл $p(T)$ не є інваріантним відносно даного марківського ланцюга. Цю ситуацію легко виправити, якщо взяти схему Метрополіса-Гастингса, де в якості апріорного розподілу фігурує розподіл (4). Зауважимо, що на відміну від схеми Гіббса, схема Метрополіса-Гастингса з апріорним розподілом (4) легко розпаралелюється і на практиці в деяких ситуаціях може працювати швидше, аніж схема Гіббса [12].

2.2 Методи кластерного аналізу

Під терміном кластеризація розуміють процес поділу множини об’єктів на групи, в кожній з яких знаходяться об’єкти схожі між собою. Ці групи називають кластерами.

Застосування кластерного аналізу включає 5 основних етапів:

1. Відбір вибірки об’єктів для кластеризації.

2. Визначення властивостей та ознак, за якими будуть оцінюватися об’єкти в вибірці.

3. Обчислення значень тієї або іншої міри подібності між об’єктами.

4. Застосування кластерного аналізу для створення груп подібних об’єктів.

5. Подання результатів аналізу.

Після отримання результатів можливе коригування обраної міри, кількості кластерів або ж взагалі методу кластеризації, для отримання оптимального результату.

Міра зазвичай обирається в залежності від простору, в якому знаходяться об’єкти або ж від неявних характеристик кластерів. Тож, розглянемо існуючі міри, за якими можна визначати подібність об’єктів:

1. Евклідова відстань − являє собою геометричну відстань в багатовимірному просторі:

$$\rho(x, x') = \sqrt{\sum_{i}^{n} (x_i - x_i')^2}$$

2. Квадрат евклідової відстані − зазвичай використовується для надання більшої ваги віддаленим об’єктам:

$$\rho(x, x') = \sum_{i}^{n} (x_i - x_i')^2$$

3. Манхетенська відстань − зменшує вплив окремих великих різниць між одноіменними координатами точок, так як при обчисленні відстані ці різниці не зводяться до квадрату. В більшості випадків приводить до таких самих результатів, що і для звичайної Евклідової відстані:

$$\rho(x, x') = \sum_{i}^{n} |x_i - x_i'|^2$$

4. Відстань Чебишева – використовується, коли треба визначити два об’єкти як «різні», якщо вони відрізняються по якій-небудь одній координаті:

$$\rho(x, x') = \max(|x_i - x_i'|)$$

5. Відстань Мінсковського − використовують коли треба збільшити чи зменшити вагу, що відноситься до розмірності, для якої відповідні об’єкти сильно видрізняються [13].

Самі методи кластеризації поділяються на ієрархічні та неієрархічні. Ієрархічні(деревоподібні) алгоритми будують не одне розбиття вибірки на кластери, що не перетинаються, а систему вкладених розбиттів, кожне з яких відповідає одному кроку алгоритму. Тобто на виході ми отримуємо дерево кластерів, корнем якого є вся вибірка, а «листям» - більш малі кластери. Методи такого типу засновані на обчисленні матриці подібності, що містить міри відстаней, тобто подібність об’єктів.

Серед ієрархічних алгоритмів виділяються два основних типа: агломеративні та диві зимні(Рисунок 2.1).

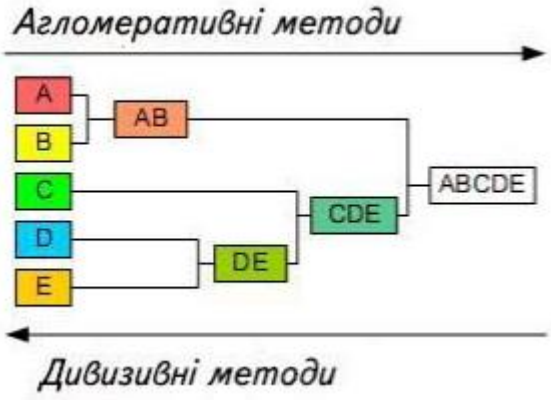

Рисунок 2.1 Принцип роботи ієрархічних алгоритмів

Методи агломеративноо типу засновані на об’єднанні об’єктів і відповідним зменшенням кількості кластерів, тобто на першому етапі кожен об’єкт знаходиться в окремому кластері. Надалі на кожному кроці роботи алгоритму відбувається об'єднання двох найближчих кластерів та переобчислення матриці відстаней, розмірність якої, очевидно, знижується на одиницю. Робота алгоритму закінчується, коли всі об’єкти стають членами одного кластеру.

Ці методи розрізняються правилами побудови кластерів, серед яких: одиничного зв’язку, повного зв’язку, середнього зв’язку та метод Варда.

Принцип роботи ієрархічних дівізівних процедур, навпаки, полягає в

послідовному поділі груп елементів. На першому етапі усі об’єкти належать одному класу, а надалі діляться на менші кластери і в результаті виходить послідовність груп.

До недоліків ієрархічних палгоритмів слід віднести громіздкість їх обчислювальної реалізації. На кожному кроці алгоритми вимагають обчислення матриці відстаней, а отже, ємною машинної пам'яті і великої кількості часу.

Неієрархічні методи являють собою ітеративні методи дроблення вихідної сукупності. В процесі розподілу формуються нові кластери, і так до тих пір, поки не буде виконано правило зупинки. Між собою методи розрізняються вибором початкової точки, правила формування нових кластерів і правилом зупинки. Найчастіше використовується алгоритм К-середніх. Він має на увазі, що аналітик заздалегідь фіксує кількість кластерів в результуючому розбитті [14].

Завдання кластеризації цим методом можна розглядати як побудову оптимального розбиття об'єктів на групи. При цьому оптимальність може бути визначена як вимога мінімізації середньоквадратичної помилки розбиття(мінімізація сумарного квадратичного відхилення точок кластерів від центрів цих кластерів): $V = \sum_{i=1}^{k} \sum_{x \in S_i}^{n_j} (x - \mu_i)^2$, де k - число кластерів, S_i - отримані кластери, $i = 1, 2, ..., k$, μ_i - центри мас усіх векторів x з кластеру S_i.

Принцип роботи даного алгоритму такий:

1. Випадково обираються k точок, що є початковими «центрами мас» кластерів.
2. Віднесення кожного об’єкту до кластера з найближчим «центром мас»
3. Переобчислення «центру мас» кластерів відповідно до поточного складу.
4. Якщо критерій зупинки алгоритму не задоволений, повернутися до кроку 2

В якості критерія зупинки обирають один з двох:

1. Відсутність переходу об’єктів з кластеру в кластер на кроці 2
2. Мінімальні зміни середньоквадратичної помилки.

Зупинка алгоритму відбувається за кінцеву кількість ітерацій, так як кількість можливих розбиттів кінцевої множини кінцева, а на кожному кроці сумарне квадратичне відхилення зменшується, тому за циклювання неможливо.

Перевагами цього методу є простота використання, швидкість використання, зрозумілість і прозорість алгоритму.

До недоліків даного алгоритму можна віднести:

- необхідність задавати кількість кластерів для розбиття;
- результат залежить від вибору початкових центрів кластерів;
- алгоритм занадто чутливий до викидів, які можуть спотворювати середнє;
- повільна робота на великих базах даних.

Існує декілька способів усунення цих недоліків:

- декілька випадкових кластеризацій;
- поступове нарощення кількості кластерів k [15].

2.2.1 Порівняння ієрархічних та неієрархічних методів кластеризації

Аби надати перевагу якомусь типу методів, при реалізації поставленої задачі, треба врахувати особливості кожного з запропонованих типів.

До переваг неієрархічних методів можна віднести вищу стійкість щодо шумів і викидів, некоректного вибору метрики, введення незначущих змінних у набір, що бере участь у кластеризації. Але при цьому треба визначати кількість кластерів, кількість ітерацій або правило зупинки, а також деякі інші параметри кластеризації. Якщо кількість кластерів важко визначити, то краще використати ієрархічні алгоритми. Якщо вибірка дуже велика, то можна провести низку експериментів з

різною кількістю кластерів, або ж поступово нарощувати кількість кластерів та порівнювати результати.

Оскільки ієрархічні методи, на відміну від неієрархічних, не потребують заданя кількості кластерів, вони будують повне дерево вкладених кластерів, тому результати виявляються досить наочними та детальними.

Та все ж мають і свої недоліки: обмеження обсягу набору даних, вибір міри подібності, негнучкість отриманих класифікацій.

Використовуючи ієрархічні методи, можливо доволі легко ідентифікувати викиди в наборі даних й, у результаті, підвищити якість даних [16, 17].

Таблиця 2.1 Порівняння алгоритмів кластеризації

Алгоритм кластеризації	**Форма кластерів**	**Вхідні дані**	**Результати**
Ієрархічний	Довільна	Кількість кластерів або поріг відстані для усікання ієрархії	Бінарне дерево кластерів
k-середніх	Гіперсфера	Кількість кластерів	Центри кластерів
c-середніх	Гіперсфера	Число кластеров, степень нечеткости	Центри кластерів, матриця приналежності
Виділення зв’язних компонент	Довільна	Поріг відстані	Деревовидна структура кластерів
Минимальное покрывающее дерево	Довільна	Кількість кластерів або поріг відстані для видалення ребер	Деревовидна структура кластерів
Пошарова кластеризація	Довільна	послідовність порогів відстані	Деревовидна структура кластерів з різними рівнями

Таблиця 2.2 Обчислювальна складність алгоритмів

Алгоритм кластеризації	**Обчислювальна складність**
Ієрархічний	$O(n^2)$
k-середніх	O(nkl), де k – число кластерів, l – число ітерацій
c-середніх	
Виділення зв'язних компонент	Залежить від алгоритму
Мінімальне кістякове дерево	$O(n^2 \log n)$
пошарова кластеризация	O(max(n, m)), где m < n(n-1)/2

Висновки до розділу 2

У даному розділі описано Байєсівський підхід та його застосування у методах Монте-Карло, з метою обчислення імовірності наявності необхідної інформації серед численних даних. Також було розглянуто різні методи кластерного аналізу, що дозволяють знайти приховані зв’язки між даними. Звернуто увагу на переваги та недоліки, розглянутих методів при обробці даних.

Розглянувши теоретичну частину цих методів стало зрозуміло, що їх використання задля виявлення взаємозв’язків між подіями в лог-файлах, може давати досить вражаючу точність розподілення даних на класи подібних об’єктів та пошуку зв’язків між даними, які є невидимими для людського ока. З практичної точки зору, застосування розглянутих алгоритмів до цієї задачі, є складно реалізованим, адже для обробки терабайтів даних необхідна вкрай велика обчислювальна потужність ЕОМ. Також однією з складнощів використання даних алгоритмів до журналів реєстрації подій є необхідність перетворення безлічі різнорідних текстових даних до числового формату.

3 МЕТОДИКА ВИЯВЛЕННЯ НАЯВНОСТІ В МЕРЕЖАХ НЕВРАХОВАНИХ ПЕРСОНАЛЬНИХ ДАНИХ

Як вже зазначалося раніше, глобальні та корпоративні мережі нілічують терабайти даних, які потребують надійного захисту від несанкціонованого використання. Ці дані представлені у багатьох форматах, кожен з яких несе в собі свою цінність. В даний час все більше уваги приділяється даним, що являють собою персональні дані суб'єктів. Такого роду інформація підпадає під особливий нагляд,− створюється все нові і нові правила та вимоги щодо забезпечення її безпеки. Разом з ними зростає різноманітність злочинних інформаційних атак на ці дані.

Під пильний нагляд злочинців підпадають журнали реєстрації подій глобальних та корпоративних мереж, адже володіючи цими даними, можна виявити досить цінну інформацію про суб'єктів користування web-ресурсами.

3.1 Виявлення персональних даних

Розглянувши технології обробки даних великих об'ємів, правила та вимоги обробки даних, зокрема персональних, а також методи їх обробки,− надалі дослідимо ситуацію з виявленням в мережах неврахованих персональних даних на основі одного з можливих способів їх отримання – дослідження журналів реєстрації подій.

Даний випадок розглянемо із застосуванням платформи Splunk, яка за допомогою методів, описаних у цій роботі, допоможе виявити приховані зв'язки між логами та скласти «портрет користувача».

Для практичної реалізації поставленої задачі було взято журнали логів Інтернет-магазину комп'ютерних ігор, які містили access- та secure-логи. Дослідивши вміст даних файлів було виявлено, що серед тисяч рядків

зареєстрованих подій, тільки частина містить інформацію, проаналізувавши яку можна отримати дані про користувачів цього магазину. В нашому випадку, під інформацією розуміємо: час перебування на сайті, ім'я користувача(user_name), місце знаходження(ip-адреса), браузер, ОС, пристрій, що використовували для перегляду сторінок сайту, перелік посилань за якими переходив користувач та інше(Рисунок 3.1).

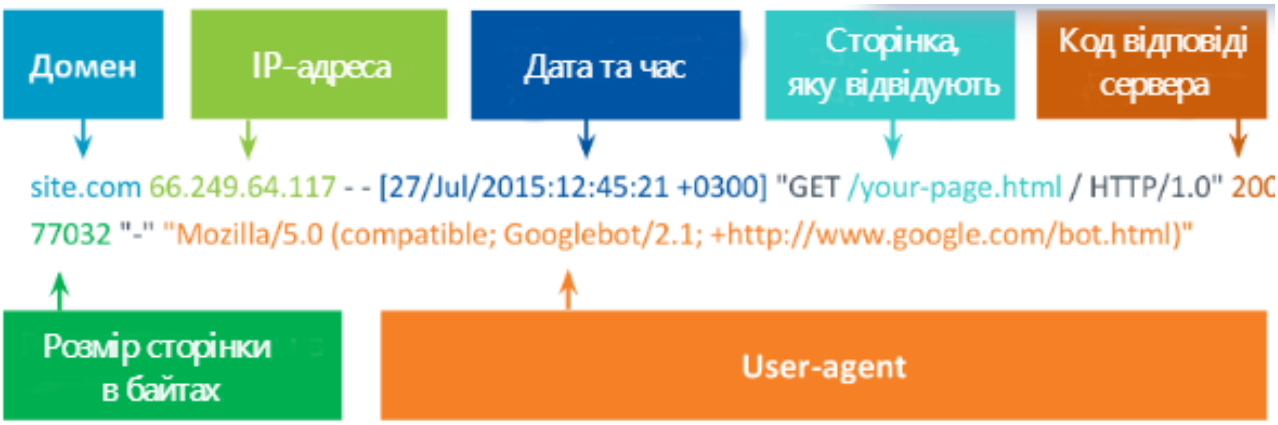

Рисунок 3.1 Структура події з лог-файлу

Розглянувши ці дані окремо один від одного − не отримаємо нічого цінного, а от якщо мати цілий «портрет користувача» з усієї цією інформацією, то тоді можна використати її задля злочинної діяльності.

Практична реалізація задачі виявлення неврахованих персональних даних була виконана за допомогою лог-файлів одного з інтернет-магазинів комп'ютерних ігор. Взявши журнали реєстрації подій за період з 28.11.18 по 5.12.18(8 днів), було виявлено 219728 подій. Виокремивши усі ip-адреси наявні в логах,було візуалізовано країни, користувачі з яких заходили на сайт цього магазину(Рисунок 3.2). Запит для візуалізації такої карти в Splunk виглядає так: sourcetype=access_combined_wcookie | iplocation clientip | stats count by Country | geom geo_countries featureIdField="Country".

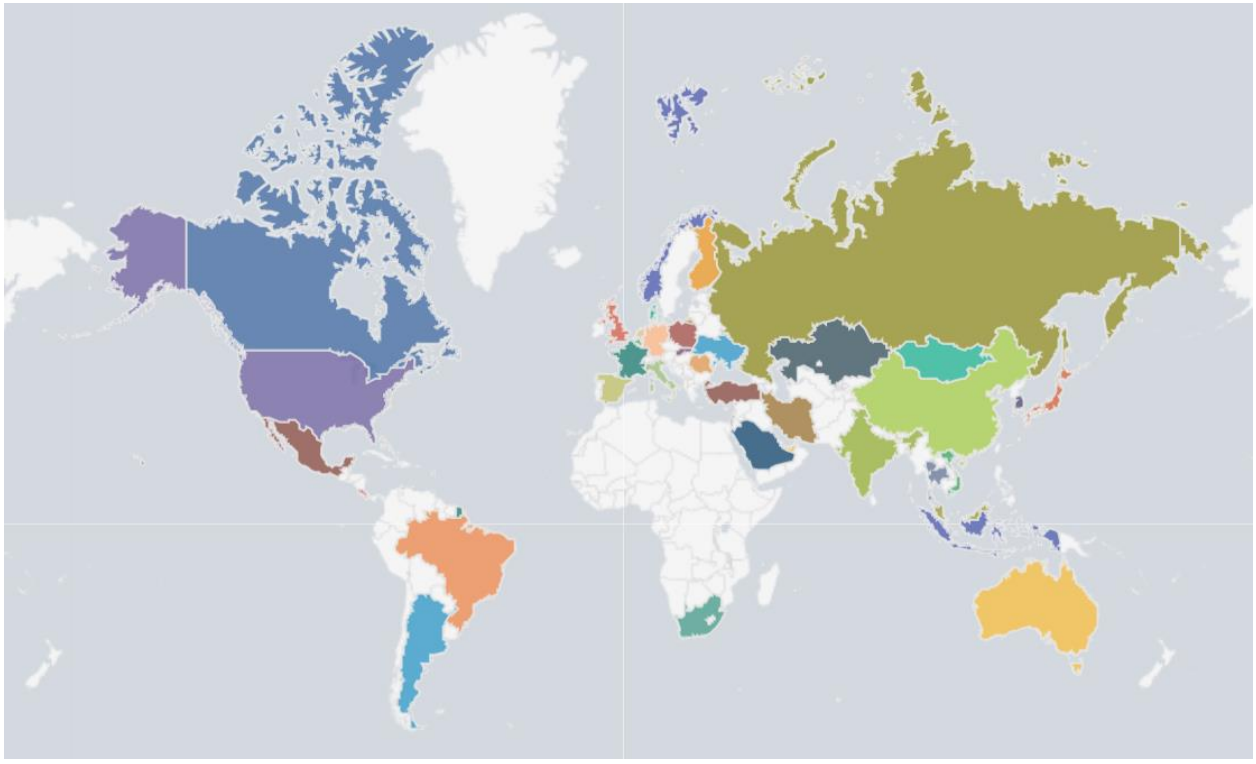

Рисунок 3.2 Країни, з яких переходили на сайт інтернет-магазину

Розберемо декілька прикладів, що наочно продемонструють наслідки неналежного ставлення до захисту персональних данних в комп'ютерних мережах:

1. Підбір паролю до особистого кабінету користувача інтернет-магазину по знайденому в журналах user_name.

Найпростіший випадок, коли у вмісті події знайдено user_name користувача інтернет-магазину (Рисунок 3.3). Маючи в руках таку інформацію, можна підібрати пароль до особистого кабінету суб'єкта та отримати повний доступ до персональних даних користувача, які він надав для зручної комунікації з продавцем продукту, наприклад, це може бути номер телефону, адреса місця проживання, пошта, можливо інформація про сімейний стан, вік, релігійну приналежність та інше. В залежності від повноти отриманої інформації можна фізично знайти людину, заподіяти шкоди web-додатку, виконати грошові махінації за кошти користувача ресурсу, надіслати на пошту користувача посилання з вірусом нібито від цього сайту інтернет-магазину у вигляді якоїсь акції або подарунку та інше.

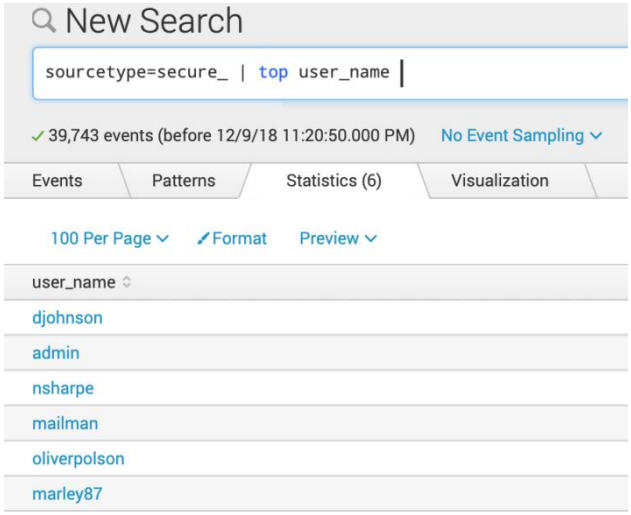

Рисунок 3.3 Список user_name

Розглянувши наявні в лог-файлах події, було виявлено досить багато імен користувачів. Деякі з них були абсолютно беззмістовні (tkp_00, 58hlo08), але й були такі, де користувач в якості user_name використовував своє ім’я та прізвище або ж першу літеру ім’я та прізвище (Рисунок 3.3). В залежності від намірів злочинця, такая інформація теж може стати ціною в його руках, не дивлячись на те, що існує імовірність того, що ім’я та прізвище − несправжні. Взявши user_name oliverpolson, ip-адресу з якої він заходив на сторінку магазину та скориставшись, наприклад, телефонними довідниками або ж соціальними мережами,− можна знайти точну адресу, номер телефону та іншу інформацію про цього користувача.

2. Стеження за діяльністю користувача інтернет-магазину

Кожен крок користувача інтернет-магазину записаний у журналах реєстрації подій, тож не виникає ніякої складності для стеження за діями суб’єкта та використання цієї інформації зловмисниками.

Оскільки ip-адреси належать не самим суб’єктам, а провайдерам, якими вони користуються, то з однієї такої адреси може заходити безліч користувачів. Такі випадки є досить складними для аналізу, але не варто виключати випадок, коли за певний період часу на web-ресурс з певною ip-адресою завітав лише один користувач. Тож, перевіривши дану гіпотезу на

наявних даних, дійсно впевнилися, що така ситуація існує (Рисунок 3.4).

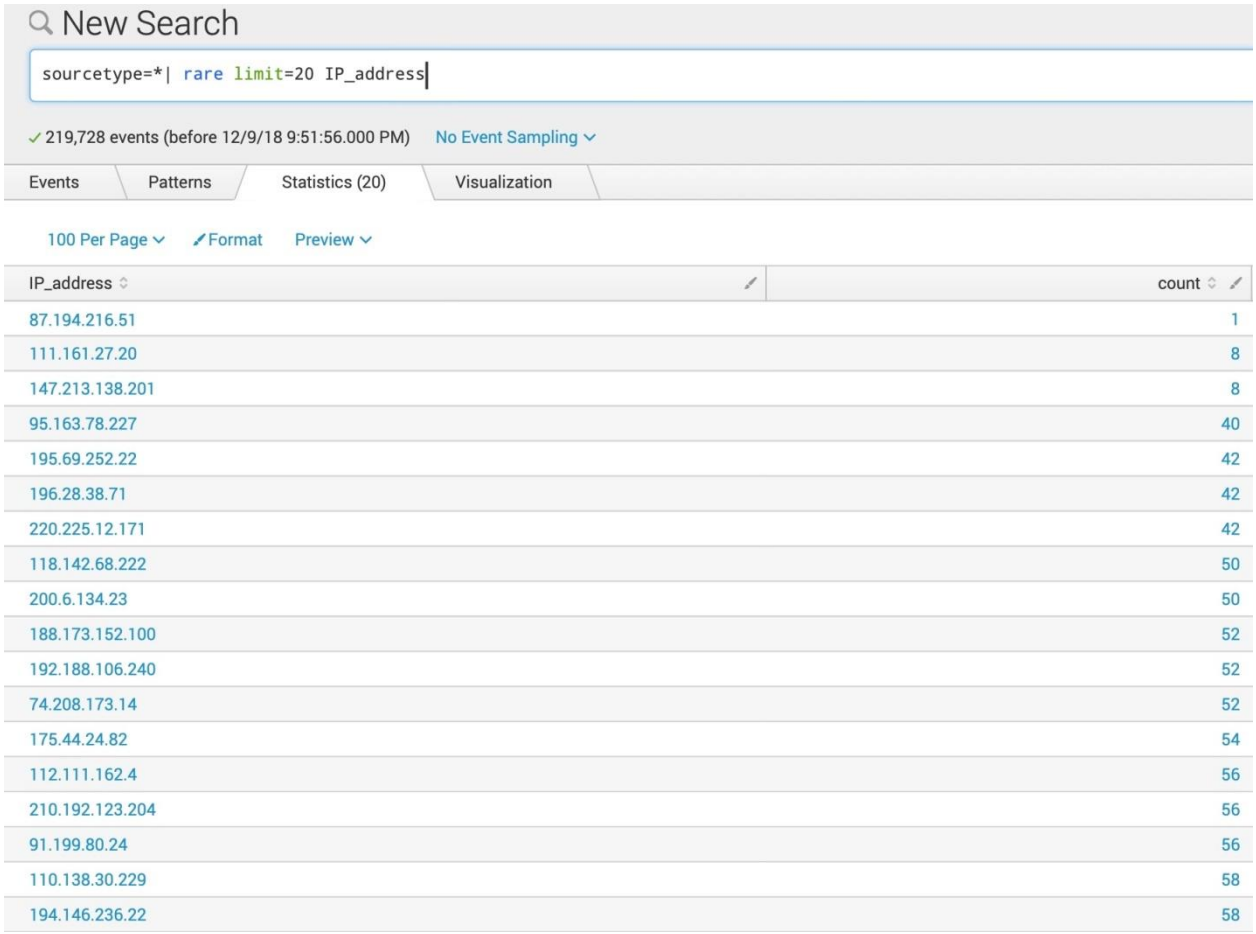

Рисунок 3.4. Відповідність кількості користувачів до ip-адрес з яких вони заходили на сайт

Тоді, спробуємо взявши цього суб'єкта, скласти «портрет користувача» зібравши усю можливу інформацію пов'язану з ним. Тож, ip-адреса 84.194.216.51 відповідає користувачу marley87.

Простеживши діяльність суб'єкта на проміжку часу, дані за який ми маємо, отримуємо такі результати:

1. Активність користувача під user_name marley87 − 1382 подій за 8 днів. Серед них:

- 11 невдалих спроб увійти у свій особистий кабінет
- здійснено 4 покупки (з productId наявних в логах, дізнаємося який саме продукт та за яку ціну був придбаний)
- даний користувач заходив на сайт переважно в ранковий час (з 6 до 12 години)

- цей користувач заходив постійно з однієї і тої самої ip-адреси

- також з логів видно, що вхід здійснювався з одного пристрою, браузеру та пошукової системи.

За допомогою команди: sourcetype=access_combined_wcookie clientip="87.194.216.51" action=addtocart OR action=purchase OR action=view OR action=remove OR action=changequantity, отримуємо інформацію щодо перегляду користувачем продуктів, додавання їх до кошика, здійснення покупки, видалення з кошика, зміну кількості доданих продуктів до кошика.

Приклади подій з вище вказаною інформацією:

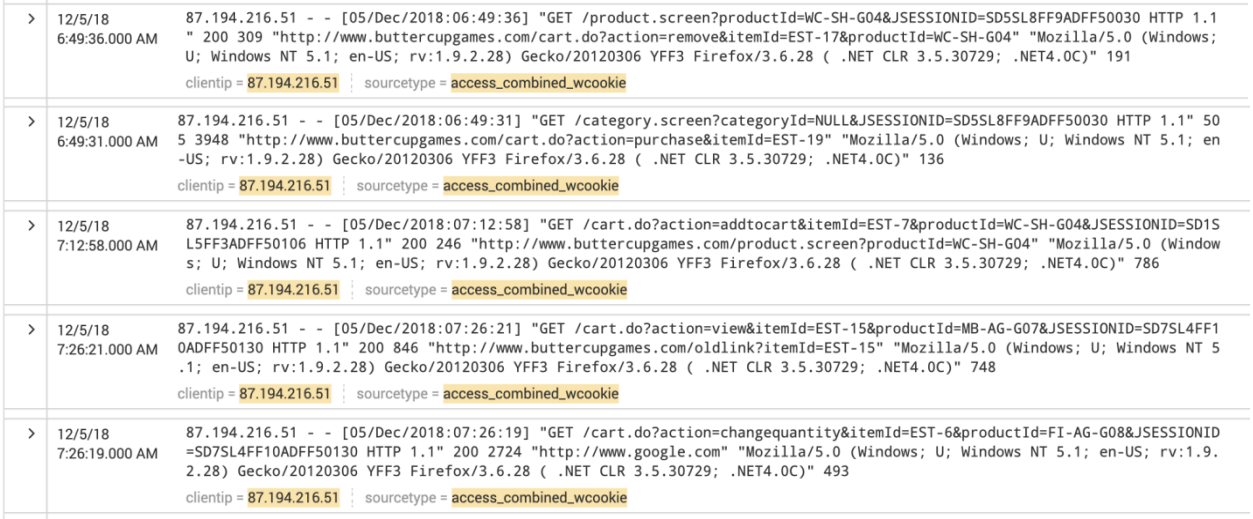

Рисунок 3.5 Access-логи

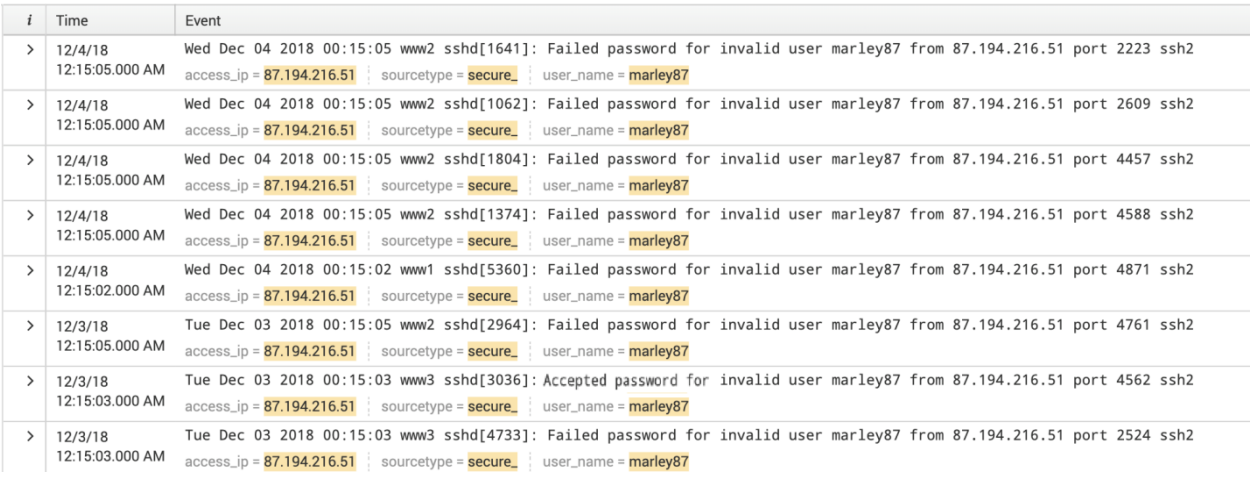

Рисунок 3.6 Secure-логи

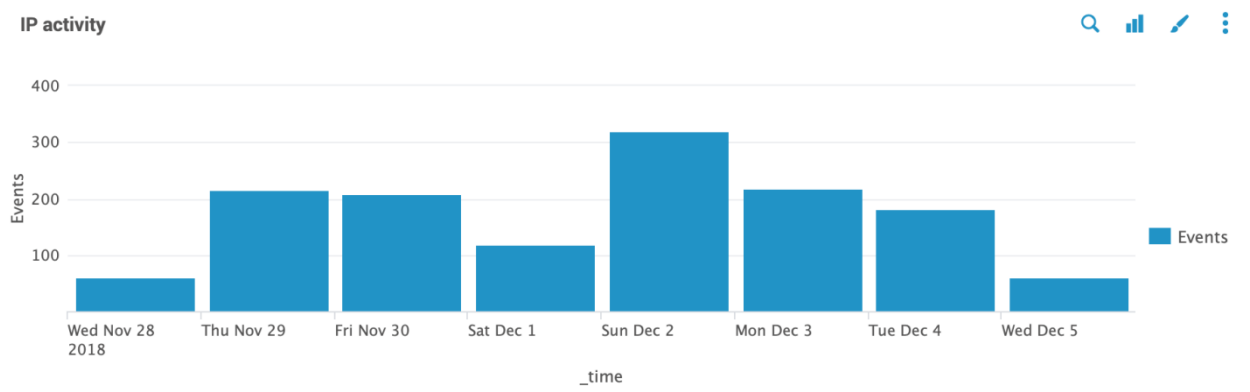

Рисунок 3.7 Загальна активність користувача marley87

Тож, отримані результати дають досить багато приватної інформації про користувача.

Як вже було сказано, в залежності від цілей злочинців, інформацію з журналів можна використовувати по-різному. Ще одним прикладом використання є дослідження кількості покупок або ж кількості переглядів продукту або сайту вцілому по країні або місту(Рисунок 3.8).

Рисунок 3.8 Активність користувачів з Пекіну

Висновки до розділу 3

В даному розділі було розглянуто реалізацію декількох прикладів з виявлення персональних даних в глобальних та локальних мережах. Було застосовано пакети агрегації і обробки до подій для виявлення кореляції між

ними: від найпростішого вивлення однакових полів в подіях до обчислення статистичних характеристик для вивлення прихованих зв'язків. Було акцентовано увагу на багато факторів, що перешкоджають зловмисникам зібрати «портрет користувача» та вчинити шахрайські дії.

На основі отриманих результатів, можна стверджувати про існування реальної можливості незаконного отримання персональних даних.

ВИСНОВКИ

Головним чином дана робота присвячена виявленню неврахованих персональних даних у глобальних та корпоративних мережах. Адже інформація такого роду, є дуже цінною річчю для злочинців, які кожен день вигадують все нові способи заволодіння чужою інформацією. Існіючі мережі містять у собі дуже великі масиви даних, і з кожним днем їх стає все більше. В свою чергу людство створює спеціальні автоматизовані системи, що допомагають їм знаходити потрібну інформацію, відмітаючи тони сміття, а також знаходити зв’язки між купою даних.

Одним найбільш цінних типів даних для інтернет-злочинців є персональні дані користувачів web-ресурсів. Рушієм для дослідження існування в глобальних та корпоративних мережах неврахованих персональних даних стала подія набуття чинності принципово нового законодавства про безпеку персональних даних(GDPR), завдяки чому підвищується актуальність задачі забезпечення захисту персональних даних.

У даній роботі розглянуто клас систем, які за допомогою свого потужного механізму дозволяють збирати, обробляти та аналізувати великі масиви даних, таких як Splunk та IBM i2. Також було досліджено вимоги та правила згідно з якими повинні використовуватися та оброблятися персональні дані. Крім того, досліджено методи кластерного аналізу та Баєсівський підхід у методах Монте-Карло, що допомагають здійснювати аналіз даних великих об'ємів. Звернуто увагу на проблеми з якими можна зіткнутися при практичній реалізації задачі вивлення персональних даних в комп'ютерних мережах: необхідність великої обчислювальної потужності ЕОМ, складність перетворення великого об'єму різнорідного тексту подій в числовий формат та інші.

На основі інформаії, що була викладена в даній роботі та практично реалізованих прикладів виявлення персональних даних у журналах реєстрації подій, була створена методика виявлення неврахованих персональних даних

в комп'ютерних мережах. Беручи до уваги дану методику, в подальшому, організації та компанії, що володіють такими даними зможуть більш ефективно їх захищати.

ПЕРЕЛІК ПОСИЛАНЬ

1. Carasso D. Exploring Splunk, 2012. vol. 2. pp. 23-29.

2. Zadrozny P., Kodali R. Big Data Analytics Using Splunk: Deriving Operational Intelligence from Social Media, Machine Data, Existing Data Warehouses, and Other Real-Time Streaming Sources. 1st ed. 2013. vol. 5. pp. 176-179.

3. Buck J., Fottral J. Mastering IBM i: The Complete Resource for Today's IBM i System. 2011. pp. 563-567.

4. Woodbury C. IBM i Security Administration and Compliance. 2016. p. 327

5. Voigt P., Bussche A. The EU General Data Protection Regulation (GDPR). 2017. pp. 214-218.

6. Положення Загального регламенту захисту даних (General Data Protection Regulation, GDPR; Regulation (EU) 2016/679) [Електронне джерело]. Режим доступу:https://www.kmu.gov.ua/storage/app/media/uploaded-files/es-2016679.pdf. Дата фіксації 10.11.18

7. Zander T. OWASP Top 10: The Top 10 Most Critical Web Application Security Threats. 2015. pp. 31-33.

8. Переклад OWASP Testing Guide. Частина 3.4. [Електронне джерело]. Режим доступу: https://defcon.ru/web-security/4249/ . Дата фіксації 10.11.18

9. Carlin B.P., Louis T.A. Bayesian Methods for Data Analysis. 3rd Edition. 2008. pp. 356-378.

10. Gelman A. Carlin J.B., Stern H.S., Rubin D.B. Bayesian Data Analysis, Second Edition. 2nd Edition. 2003. pp. 553-587.

11. Gamerman D. Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference. 1997. pp. 152-154.

12. Ветров Д.П. Методы Монте Карло по схеме марковских цепей. 2017. –9 с.

13. Хотинский А.М., Королёва С.Б. Факторный, дискриминантный и кластерный анализ. –М.: Финансы и статистика, 1989. – 216 с.

14. Дубровская Л.И., Князев Г.Б. Компьютерная обработка естественно-научных данных методами многомерной прикладной статистики: Учебное пособие. – Томск: ТМЛ-Пресс, 2011. –120 с.

15. Coates A., Andrew Y. Ng. Learning Feature Representations with K-means. Stanford University, Stanford, USA. 2012. pp. 2-5.

16. Everitt B.S., Landau S., Leese M. Cluster analysis. 2001. pp. 243-256

17. Aggarwal C.C., Reddy C.K. Data Clustering: Algorithms and Applications 1st Edition. 2013. pp. 458-473.